Biometrics 61, 186–192
March 2005

Genomic Control for Association Studies under Various Genetic Models

Gang Zheng,[1,*] Boris Freidlin,[2,] Zhaohai Li,[3,4,***] and Joseph L. Gastwirth[3,4,****]**

[1]Office of Biostatistics Research, DECA, National Heart, Lung and Blood Institute, 6701 Rockledge Drive, MSC 7938, Bethesda, Maryland 20892-7938, U.S.A.
[2]Biometric Research Branch, DCTD, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.
[3]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.
[4]Department of Statistics, George Washington University, Washington, D.C. 20052, U.S.A.
[*]*email:* zhengg@nhlbi.nih.gov
[**]*email:* freidlinb@ctep.nci.nih.gov
[***]*email:* zli@gwu.edu
[****]*email:* jlgast@gwu.edu

Summary. Case–control studies are commonly used to study whether a candidate allele and a disease are associated. However, spurious association can arise due to population substructure or cryptic relatedness, which cause the variance of the trend test to increase. Devlin and Roeder derived the appropriate variance inflation factor (VIF) for the trend test and proposed a novel genomic control (GC) approach to estimate VIF and adjust the test statistic. Their results were derived assuming an additive genetic model and the corresponding VIF is independent of the candidate allele frequency. We determine the appropriate VIFs for recessive and dominant models. Unlike the additive test, the VIFs for the optimal tests for these two models depend on the candidate allele frequency. Simulation results show that, when the null loci used to estimate the VIF have allele frequencies similar to that of the candidate gene, the GC tests derived for recessive and dominant models remain optimal. When the underlying genetic model is unknown or the null loci and candidate gene have quite different allele frequencies, the GC tests derived for the recessive or dominant models cannot be used while the GC test derived for the additive model can be.

Key words: Cochran–Armitage trend test; Genetic model; Optimal genomic control; Variance inflation factor.

1. Introduction

Case–control studies testing linkage disequilibrium or association provide a more powerful method than linkage studies for detecting small genetic effects on traits (Risch and Merikangas, 1996). One drawback of the case–control design is that it may produce spurious association due to population substructure. Substructure also creates departure from Hardy–Weinberg equilibrium (HWE) that induces a positive dependence in the alleles in randomly selected individuals. On the other hand, family-based association studies, such as the case–parents trio design (Spielman, McGinnis, and Ewens, 1993; Schaid and Sommer, 1993), use family members as controls and reduce the effect of population substructure. To carry out a family-based study genotypes of both cases and their family members, including parents, are obtained. This is often difficult for studies of late onset diseases. In case–control studies, however, population controls are used as parental genotypes are not needed. Moreover, it is shown that case–control studies are more powerful than family-based association studies (Risch and Teng, 1998). If the hidden population stratification can be adjusted, it is much easier to recruit population controls than family-based controls. Recently, several methods have been proposed to adjust the effect of population substructure. The genomic control (GC) of Devlin and Roeder (1999) is an important one of them, which is also the focus of this article.

To adjust for hidden population stratification in case–control studies, Devlin and Roeder (1999) proposed a novel GC method using trend tests (see also Devlin, Roeder, and Wasserman, 2001). When the hidden population stratifications exist, the variance of the trend test would be inflated. The principle of GC method is that, if the population stratification affects the candidate gene, it will also affect the unrelated null loci. Hence, the variance inflation of the trend test for the null loci can be used to adjust that of the trend test for the candidate gene. The key assumption of the GC approach is that, under the null hypothesis of no association, the trend test statistics for unrelated genes are independent and follow the same distribution. So the variance inflation is a constant across unrelated genes on the same chromosome.

Under this assumption, they derived a variance inflation factor (VIF) for the trend test to estimate the VIF using null loci and the variance of the trend test for the candidate gene is then adjusted. The GC tests have been shown more powerful than some family-based tests using trios (Bacanu, Devlin, and Roeder, 2000).

In genetic analysis, a genetic model refers to the mode of inheritance corresponding to a recessive, additive, or dominant disease. For complex diseases, however, the underlying genetic model is usually unknown. In that situation, the additive genetic model is easy to use and it has some robustness properties, especially when the recessive model can be eliminated (Freidlin et al., 2002). The GC test given in Devlin and Roeder (1999) was derived assuming an additive genetic model. But the properties of GC under other genetic models have not been reported. In this article, we examine its properties for recessive and dominant models. In Section 2, the GC of Devlin and Roeder (1999) is reviewed. The VIFs for recessive and dominant models are obtained in Section 3. Simulation studies comparing GC under various genetic models are reported in Section 4. Discussion of using GC when the genetic model is unknown is given at the end.

2. Review of the GC Test

In case–control studies, cases and controls are sampled independently and their genotypes are obtained. Let A be the high-risk candidate allele and a the normal one. The data from a case–control design are summarized in a 2×3 table (see Table 1), where $r_i(s_i)$, $i = 0, 1, 2$ is the number of cases (controls) whose genotypes have i A alleles.

Denote the penetrances as $f_i = \Pr(\text{case} \mid iA \text{ alleles in genotype})$, $i = 0, 1, 2$. Under the null hypothesis of no association, $f_0 = f_1 = f_2$, and under the alternative hypothesis, $f_0 \le f_1 \le f_2$, where at least one inequality holds. The recessive, additive, and dominant genetic models are defined by $f_0 = f_1$, $f_1 = (f_0 + f_2)/2$, and $f_1 = f_2$, respectively. Sasieni (1997) suggested using Cochran–Armitage (CA) trend test (Cochran, 1954; Armitage, 1955) for data given in Table 1. When the scores $x = (x_0, x_1, x_2)$, $x_0 \le x_1 \le x_2$, are assigned to three genotypes (aa, Aa, AA), respectively, the CA trend test can be written as

$$Z^2 = \frac{N\left\{\sum_{i=0}^{2} x_i(Sr_i - Rs_i)\right\}^2}{RS\left\{N\sum_{i=0}^{2} x_i^2 n_i - \left(\sum_{i=0}^{2} x_i n_i\right)^2\right\}}. \tag{1}$$

Note that the CA trend test is invariant to a linear transformation of the scores x, so all scores are of form $x = (0, \eta, 1)$, $0 \le \eta \le 1$. For any η, Z^2 in (1) is denoted by Z_η^2. Note that Z_0^2, $Z_{1/2}^2$, and Z_1^2 are the CA trend tests based on optimal scores for the recessive, additive, and dominant models, respectively (Sasieni, 1997; Freidlin et al., 2002; Zheng et al., 2003).

Table 1
Genotype distribution

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Following Devlin and Roeder (1999), let $G_i^{\text{add}} = 0, 1, 2$, $i = 1, \ldots, R$, respectively, when the ith case has genotype aa, Aa, AA. Similarly, H_j^{add}, $j = 1, \ldots, S$, are defined for controls. Consider the test statistic defined by the difference in the number of A alleles between cases and controls, $T^{\text{add}} = \sum_i G_i^{\text{add}} - \sum_j H_j^{\text{add}}$. When $R = S$, the trend test $Z_{1/2}$ is proportional to T^{add}. Assume these samples came from m subpopulations, which contribute $a_1, \ldots, a_m$ cases and $b_1, \ldots, b_m$ controls, respectively. Thus, $\sum_k a_k = R$ and $\sum_k b_k = S$. Moreover, assume that genotypes of members of the different subpopulations are independent.

Devlin and Roeder defined the VIF (λ) relative to the trend test $Z_{1/2}$ as $\lambda_{1/2} = \text{var}_{H_0}(T^{\text{add}})/\{4Rpq(1+F)\}$, where p is the frequency of allele A in the population, $q = 1 - p$, and F is Wright's coefficient of inbreeding. In the following all variances and covariances are evaluated under the null, so the subscript H_0 is omitted. Note that the denominator, $4Rpq(1+F)$, is the asymptotic variance of the corresponding score function with $x = (0, 1, 2)$ under the null hypothesis, as $N \to \infty$ and m bounded, i.e., $\text{var}(\sum_i x_i r_i - \sum_i x_i s_i) \approx 4Rpq(1+F)$. Generally, when $R \ne S$, $T^{\text{add}} = S\sum_i G_i^{\text{add}} - R\sum_j H_j^{\text{add}}$ should be used and the denominator of $\lambda_{1/2}$ is the asymptotic variance of $S\sum_i x_i r_i - R\sum_i x_i s_i$. Assuming $\text{var}(G_i^{\text{add}}) = \text{var}(H_j^{\text{add}})$ and $\text{cov}(G_i^{\text{add}}, G_l^{\text{add}}) = \text{cov}(H_j^{\text{add}}, H_l^{\text{add}}) = \text{cov}(G_i^{\text{add}}, H_j^{\text{add}})$, $i \ne l$ and $j \ne l$, where subjects i, j, l are from the same subpopulation, they obtained for $R = S$

$$\begin{aligned}\text{var}(T^{\text{add}}) = {} & 2R\text{var}\left(G_1^{\text{add}}\right) \\ & + \sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_k b_k\} \\ & \times \text{cov}\left(G_1^{\text{add}}, G_2^{\text{add}}\right),\end{aligned} \tag{2}$$

where $\text{var}(G_1^{\text{add}}) = 2pq(1+F)$ and $\text{cov}(G_1^{\text{add}}, G_2^{\text{add}}) = 4Fpq$, which follow from the additive property of G_i^{add} and H_j^{add}. Thus,

$$\lambda_{1/2} = 1 + \frac{F\sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_k b_k\}}{R(1+F)}.$$

Note that $\lambda_{1/2}$ is independent of p or q but depends on a_k, b_k, m, and F. Devlin and Roeder examined the effect of population substructure on $\lambda_{1/2}$. When there is no population substructure or cryptic relatedness ($F = 0$), under the null hypothesis, $Z_{1/2}^2$ follows central chi-square distribution with one degree of freedom χ_1^2. To adjust for population substructure or cryptic relatedness, they proposed a new test statistic $Z_{1/2}^2/\lambda_{1/2} \sim \chi_1^2$ for the null model to allow for extra variance. In practice, $\lambda_{1/2}$ is estimated using data from unrelated markers (null loci). These null loci are assumed to be unrelated to the trait and segregate independently, and the effect of population substructure on them is similar to that at the trait locus.

3. VIFs for the Recessive and Dominant Models

When the underlying genetic model is recessive or dominant, it is reasonable to use Z_0^2/λ_0 and Z_1^2/λ_1, respectively, where λ_0 and λ_1 are VIFs corresponding to recessive and dominant models. As in Devlin and Roeder, the following derivations are for the situation $R = S$. The results can readily be generalized to the case when $R \neq S$.

Under the null hypothesis, it can be shown that, for recessive and dominant models, the asymptotic variances in the denominators of λ_0 and λ_1 are

$$\mathrm{var}(r_2 - s_2) \approx 2Rpq(p + qF)\{1 + p(1 - F)\},$$

$$\mathrm{var}((r_1 + r_2) - (s_1 + s_2)) \approx 2Rpq(q + pF)\{1 + q(1 - F)\},$$

respectively. Since the dominant and recessive models are symmetric, we focus on the recessive model. For the ith case, $i = 1, \ldots, R$, let $G_i^{\mathrm{rec}} = 0$ for the genotype aa or Aa and $G_i^{\mathrm{rec}} = 1$ for genotype AA. Similarly, H_j^{rec}, $j = 1, \ldots, S$, are defined for controls. When $R = S$, Z_0 is proportional to $T^{\mathrm{rec}} = \sum_i G_i^{\mathrm{rec}} - \sum_j H_j^{\mathrm{rec}}$. To evaluate $\mathrm{var}(T^{\mathrm{rec}})$, we need to calculate $\mathrm{var}(G_1^{\mathrm{rec}})$ and $\mathrm{cov}(G_1^{\mathrm{rec}}, G_2^{\mathrm{rec}})$. As in Devlin and Roeder (1999), it can be shown that $\mathrm{var}(G_1^{\mathrm{rec}}) = pq(p + qF) \times \{1 + p(1 - F)\}$. To obtain $\mathrm{cov}(G_1^{\mathrm{rec}}, G_2^{\mathrm{rec}})$, however, one requires the joint distribution of $(G_1^{\mathrm{rec}}, G_2^{\mathrm{rec}})$, i.e., $\Pr(AA, AA)$, where both members are from the same subpopulation. This differs from the derivation for the additive model as the joint distribution of G_1^{add} and G_2^{add} is not required to obtain their covariance.

Let X and Y be any two unrelated individuals, with alleles A, B and C, D, respectively, randomly selected from the same subpopulation, in which random mating occurs. Thus, the identical-by-descent (IBD) relationship among the four alleles does not depend on the arrangement of alleles within individuals, e.g., the probability that A and B are IBD equals the probability that A and C are IBD (Evett and Weir, 1998, Chapter 4).

Following Evett and Weir (1998), let θ, γ, and δ be the respective probability that any two, three, and four alleles, selected at random from the same subpopulation, are IBD, and let Δ be the probability that any two pairs of alleles, selected at random from the same subpopulation, are IBD. Here, θ is F of Section 2. Assume that evolutionary equilibrium holds, i.e., the four probabilities (θ, γ, δ, and Δ) are not changing over time. Then $\gamma = 2\theta^2/(1 + \theta)$, $\delta = 6\theta^3/\{(1 + \theta)\ (1 + 2\theta)\}$, and $\Delta = \theta^2(1 + 5\theta)/\{(1 + \theta)\ (1 + 2\theta)\}$. Let δ_0, δ_2, δ_3, and δ_4 be respective probability that the number of alleles IBD among A, B, C, D is 0, 2, 3, 4, and let δ_{22} be the probability that any two pairs of alleles among A, B, C, D are IBD. Then $\delta_0 = 1 - 6\theta + 8\gamma + 3\Delta - 6\delta$, $\delta_2 = \theta - 2\gamma - \delta + 2\delta$, $\delta_3 = \gamma - \delta$, $\delta_{22} = \Delta - \delta$, and $\delta_4 = \delta$. Substituting γ, δ, Δ as functions of θ, we obtain $\delta_0 = (1 - \theta)^3/g(\theta)$, $\delta_2 = \theta(1 - \theta)^2/g(\theta)$, $\delta_3 = 2\theta^2(1 - \theta)/g(\theta)$, $\delta_{22} = \theta^2(1 - \theta)/g(\theta)$, $\delta_4 = 6\theta^3/g(\theta)$, where $g(\theta) = (1 + \theta)\ (1 + 2\theta)$ and $\theta = F$.

The joint distribution of genotypes of two members from the same subpopulation can be expressed in terms of the five probabilities (δ_0, δ_2, δ_3, δ_{22}, δ_4) and can be reduced to a single parameter F or θ. These joint distributions are given in Table 2. An example of the derivation of $\Pr(AA, AA)$ is given in Appendix A. Some known results are readily obtained from Table 2, for example, $\Pr(AA) = \Pr(AA, AA) + 2\Pr(AA, Aa) + \Pr(AA, aa) = pF + (1 - F)p^2$ and $\mathrm{cov}(G_1^{\mathrm{add}}, G_2^{\mathrm{add}}) = 4\Pr(AA, AA) + 2\Pr(AA, Aa) + 2\Pr(Aa, AA) + \Pr(Aa, Aa) - 4p^2 = 4Fpq$.

Table 2
Twelve of sixteen joint distributions of two genotypes from the same subpopulation (G_i = Genotype i, $i = 1, 2$)

(G_1, G_2)	$\Pr(G_1, G_2)$
(AA, AA)	$\frac{6F^3p + 11F^2(1-F)p^2 + 6F(1-F)^2p^3 + (1-F)^3p^4}{(1+F)(1+2F)}$
(AA, Aa)	$\frac{4F^2(1-F)pq + 6F(1-F)^2p^2q + 2(1-F)^3p^3q}{(1+F)(1+2F)}$
(AA, aa)	$\frac{F(1-F)pq + 2(1-F)^3p^2q^2}{(1+F)(1+2F)}$
(Aa, AA)	$\Pr(AA, Aa)$
(Aa, Aa)	$4\Pr(AA, aa)$
(Aa, aa)	$\frac{4F^2(1-F)pq + 6F(1-F)^2pq^2 + 2(1-F)^3pq^3}{(1+F)(1+2F)}$
(aA, AA)	$\Pr(AA, Aa)$
(aA, Aa)	$4\Pr(AA, aa)$
(aA, aa)	$\Pr(Aa, aa)$
(aa, AA)	$\Pr(AA, aa)$
(aa, Aa)	$\Pr(Aa, aa)$
(aa, aa)	$\frac{6F^3q + 11F^2(1-F)q^2 + 6F(1-F)^2q^3 + (1-F)^3q^4}{(1+F)(1+2F)}$

In the recessive case, $\mathrm{cov}(G_1^{\mathrm{rec}}, G_2^{\mathrm{rec}}) = \Pr(AA, AA) - \Pr^2(AA)$. Using the results in Table 2 yields

$$\mathrm{cov}\left(G_1^{\mathrm{rec}}, G_2^{\mathrm{rec}}\right) = \frac{2pqF\{2p^2 + p(5 - 3p)F + (3 - 4p)F^2 - pqF^3\}}{(1 + F)(1 + 2F)}. \quad (3)$$

To evaluate $\mathrm{cov}(G_1^{\mathrm{dom}}, G_2^{\mathrm{dom}})$ for the dominant model, where $G^{\mathrm{dom}} = 0$ for aa and 1 for Aa or AA, the joint probabilities for pairs (aa, Aa) and (Aa, Aa) are also required. Using Table 2 one finds that, by interchanging p and q, $\mathrm{cov}(G_1^{\mathrm{dom}}, G_2^{\mathrm{dom}})$ is the same as (3).

From (3), the variance of T^{rec} can be written as (Appendix B)

$$\begin{aligned}\mathrm{var}(T^{\mathrm{rec}}) &= 2R\mathrm{var}\left(G_1^{\mathrm{rec}}\right) \\ &\quad + \sum_k \{a_k(a_k - 1) + b_k(b_k - 1) - 2a_kb_k\} \\ &\quad \times \mathrm{cov}\left(G_1^{\mathrm{rec}}, G_1^{\mathrm{rec}}\right) \\ &= \frac{2Rp^2q(1 + p)}{(1 + F)(1 + 2F)} + \frac{2RpqF}{(1 + F)(1 + 2F)}h_1(F, p) \\ &\quad + \frac{2pqF}{(1 + F)(1 + 2F)}h_2(F, p)\sum_k (a_k - b_k)^2,\end{aligned}$$

where $h_1(F, p) = F^2(5p - 4 - p^2) + F(3 + 3p^2 - 9p) + (1 - 3p^2 + 3p)$ and $h_2(F, p) = 2p^2 + p(5 - 3p)F + (3 - 4p)F^2 - p(1 - p)F^3$. Note that $\mathrm{var}(T^{\mathrm{rec}})$ has the same form as $\mathrm{var}(T^{\mathrm{add}})$ given by (2). By calculus, it can be shown that $h_2(F, p) \geq 0$ for any $0 \leq p, F \leq 1$. Hence, if the sizes of cases and controls in subpopulations, (a_k, b_k), $k = 1, \ldots, m$, maximize (minimize) $\mathrm{var}(T^{\mathrm{add}})$, the same (a_k, b_k), $k = 1, \ldots, m$ also maximize (minimize) $\mathrm{var}(T^{\mathrm{rec}})$.

The VIF for the trend test Z_0, appropriate for the recessive model, is

$$\lambda_0 = \frac{\mathrm{var}(T^{\mathrm{rec}})}{2Rpq(p+qF)\{1+p(1-F)\}}.$$

Interchanging p and q yields the VIF λ_1 for the trend test Z_1, appropriate for the dominant model. To examine the effect of population substructure on λ_0 and λ_1, we consider the same range of values of $F = 0.01(0.05)$, $m = 10$, and $a_k = b_l = 16$ for $k = 1, \ldots, 5$ and $l = 6, \ldots, 10$, and $a_k = b_l = 4$ for $k = 6, \ldots, 10$ and $l = 1, \ldots, 5$ considered by Devlin and Roeder (1999). Compared with $\lambda_{1/2} = 1.3$ ($F = 0.05$) and 1.06 ($F = 0.01$) reported in Devlin and Roeder, we find that λ_0 is smaller (greater) than $\lambda_{1/2}$ when $p < 0.35$ ($p \geq 0.35$) and λ_1 is larger than $\lambda_{1/2}$ for all p. For example, $\lambda_0 = 1.13(1.01)$ when $p = 0.05$ and $F = 0.05(0.01)$ and $\lambda_0 = 1.39(1.08)$ when $p = 0.45$ and $F = 0.05(0.01)$, while λ_1 ranges from 1.43 to 1.56 (1.09–1.12) for $F = 0.05(0.01)$. Further, numerical results show that λ_0 increases with p while λ_1 decreases with p. From the formulae for $\lambda_{1/2}$ and λ_0, one sees that the VIF will increase with the sample size as the $\{a_k\}$ and $\{b_k\}$ will increase.

4. Simulation Results

In order to utilize the trend test with the correct variance, one needs to estimate the VIF. One important finding in Section 3 is that VIFs for the recessive (λ_0) model and the dominant (λ_1) model are functions of the candidate allele frequency p, while the VIF for the additive ($\lambda_{1/2}$) model is independent of p. When estimating VIFs using null loci, λ_0 and λ_1 depend on the allele frequencies of null loci unless the allele frequencies of the candidate allele and null loci are close. In other words, $Z_{1/2}^2/\lambda_{1/2}$ is not affected by frequencies of the candidate gene and null loci, but both Z_0^2/λ_0 and Z_1^2/λ_1 do depend on these frequencies.

In the first simulation, we compared the type I error and power of $Z_0/\lambda_0^{1/2}$, $Z_{1/2}/\lambda_{1/2}^{1/2}$, and $Z_1/\lambda_1^{1/2}$ for recessive (REC), additive (ADD), and dominant (DOM) models, respectively. In this simulation, we assume that the candidate gene and null loci have the same allele frequency. One-sided alternatives are considered here as the high-risk allele is assumed known. Our simulation is similar to that of Devlin and Roeder (1999) who assumed that each subpopulation is in HWE. We specify p, F, the penetrances f_0, f_1, f_2, and the sample sizes for the subpopulations (a_k, b_k), $k = 1, \ldots, m$, and c the number of null loci used to estimate the VIFs. In step 1, the allele frequency p_k was generated for the kth subpopulation from the Beta distribution Beta($(1-F)p/F$, $(1-F)q/F$) for $k = 1, \ldots, m$. In step 2, for individuals in the kth subpopulation, two alleles were drawn at random from the binomial bin(2, p_k) to create a genotype at the candidate allele locus. Disease status was randomly generated depending on the number i of candidate alleles in the genotype using a Bernoulli random variable with probability f_i. The process continued until all a_k cases and b_k controls were obtained. In step 3, genotypes were generated for the c null loci using bin(2, p_k), $k = 1, \ldots, m$. Statistics $Z_{0,j}$, $Z_{1/2,j}$, and $Z_{1,j}$ at the jth locus ($j = 1, \ldots, c$) were calculated and each $\lambda_\eta^{1/2}$, $\eta = 0$, 1/2, 1, was estimated by $\lambda_\eta^{1/2} = \mathrm{median}(Z_{\eta,1}, \ldots, Z_{\eta,c})/0.456$ as in Devlin and Roeder. Then the statistics Z_0, $Z_{1/2}$, and Z_1 were adjusted by $\lambda_0^{1/2}$, $\lambda_{1/2}^{1/2}$, and $\lambda_1^{1/2}$, respectively.

To illustrate, we first considered two extreme subpopulations with 200 cases and 200 controls from each subpopulation. The simulation results are reported in Table 3 (panel A). We also consider less extreme subpopulations with 750 cases and 250 controls from one subpopulation and 250 cases and 750 controls from the other. The simulation results are reported in Table 3 (panel B). Type I error and power were based on 100,000 and 10,000 replications, respectively. The alternatives (f_0, f_1, f_2) are chosen such that the most powerful adjusted trend test has about 80% power. When $F = 0$, i.e., there is no population substructure, the adjusted type I errors typically are slightly greater than the 0.05 nominal level. This is due to the estimation of the VIFs. The only exception occurred in the rare recessive situation (0.059) and this may be a result of the small number of cases used in Table 3 (panel A). To check this, a simulation with 500 cases and 500 controls per subpopulation ($a_1 = b_2 = 500$, $a_2 = b_1 = 0$) was carried. The adjusted and unadjusted sizes for the recessive model become 0.053 and 0.051, respectively. When the allele frequency is small ($p \leq 0.1$), $Z_1/\lambda_1^{1/2}$ and $Z_{1/2}/\lambda_{1/2}^{1/2}$ have similar power properties as most cases will have only one allele. For common variants, however, there is a noticeable loss of power when $Z_{1/2}/\lambda_{1/2}^{1/2}$ is used and the trait follows a dominant or recessive pattern.

The second simulation is similar to the first one, but the candidate gene and null loci now have different frequencies. In this simulation, we only examine the effect of allele frequency discrepancy on the adjusted type I error. Results are reported in Table 4 (null loci have different allele frequencies from the candidate allele) and Table 5 (null loci have allele frequencies near that of the candidate allele). The frequencies of null loci are randomly chosen from a given uniform distribution. For example, when the candidate gene has frequency $p = 0.1$, the frequencies of null loci follow the uniform distribution (0.4, 0.6). We find that the adjusted size of the GC test for the additive model is close to the nominal level in both Tables 4 and 5 while, for the other two models, the adjusted sizes are close to the nominal level when the frequencies of null loci are close to that of the candidate gene (Table 5). The difference between the nominal and adjusted levels, however, becomes larger when the frequencies of the null loci differ from that of the candidate allele (Table 4). The empirical power is also reported in Table 5 for $F = 0.005$ when the frequencies of the null loci are close to that of the candidate gene. It shows that in this situation the GC trend tests for the dominant and recessive models can still be used except in the rare recessive situation. The power properties of three GC trend tests in Table 5 are similar to those reported in Table 3 (panels A and B).

5. Discussion

VIFs for the optimal trend tests for the recessive and dominant models are obtained here. While they have the same form as the factor for the additive model, they are functions of the frequency of the candidate allele. Simulation results demonstrate that the optimal GC tests for the recessive or dominant model can be used when the candidate gene and null loci have similar allele frequency. Because the Centers for Disease

Table 3

Type I error (adjusted and unadjusted) and power comparison of GC trend tests for three different models: candidate gene and null loci have the same allele frequency (p) and the number of null loci $c = 41$ with 10,000 *(power) and* 100,000 *(type I error) simulations*

F	p	Model	DOM	ADD	REC
		Panel A: $a_1 = 200$, $a_2 = 0$ and $b_1 = 0$, $b_2 = 200$			
0	0.1	Null	0.058[1] (0.051[2])	0.058 (0.052)	0.055 (0.059)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.167$)	0.784	0.762	0.115
		ADD ($f_0 = 0.1$, $f_1 = 0.165$, $f_2 = 0.23$)	0.792	0.801	0.218
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.475$)	0.210	0.435	0.788
	0.5	Null	0.057 (0.051)	0.056 (0.051)	0.056 (0.050)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.172$)	0.800	0.627	0.199
		ADD ($f_0 = 0.1$, $f_1 = 0.143$, $f_2 = 0.186$)	0.674	0.791	0.611
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.163$)	0.241	0.693	0.823
0.005	0.1	Null	0.057 (0.17)	0.057 (0.18)	0.036 (0.09)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.24$)	0.809	0.775	0.143
		ADD ($f_0 = 0.1$, $f_1 = 0.23$, $f_2 = 0.36$)	0.782	0.768	0.315
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.6$)	0.178	0.300	0.803
	0.2	Null	0.056 (0.16)	0.056 (0.17)	0.053 (0.10)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.205$)	0.780	0.705	0.217
		ADD ($f_0 = 0.1$, $f_1 = 0.2$, $f_2 = 0.3$)	0.819	0.802	0.523
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.32$)	0.202	0.390	0.785
	0.5	Null	0.056 (0.16)	0.056 (0.17)	0.056 (0.16)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.225$)	0.807	0.507	0.194
		ADD ($f_0 = 0.1$, $f_1 = 0.195$, $f_2 = 0.29$)	0.800	0.807	0.693
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.2$)	0.235	0.557	0.789
0.05	0.2	Null	0.050 (0.35)	0.046 (0.36)	0.044 (0.27)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.55$)	0.783	0.675	0.201
		ADD ($f_0 = 0.1$, $f_1 = 0.475$, $f_2 = 0.85$)	0.745	0.719	0.520
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.88$)	0.248	0.384	0.809
	0.5	Null	0.051 (0.34)	0.049 (0.37)	0.051 (0.34)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.65$)	0.833	0.502	0.184
		ADD ($f_0 = 0.1$, $f_1 = 0.45$, $f_2 = 0.8$)	0.729	0.742	0.677
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.45$)	0.210	0.422	0.744
		Panel B: $a_1 = 750$, $a_2 = 250$ and $b_1 = 250$, $b_2 = 750$			
0	0.1	Null	0.056[1] (0.048)[2]	0.057 (0.049)	0.054 (0.047)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.128$)	0.795	0.776	0.134
		ADD ($f_0 = 0.1$, $f_1 = 0.127$, $f_2 = 0.154$)	0.802	0.818	0.273
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.226$)	0.167	0.352	0.794
0.005	0.1	Null	0.053 (0.18)	0.053 (0.19)	0.056 (0.08)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.158$)	0.810	0.780	0.202
		ADD ($f_0 = 0.1$, $f_1 = 0.155$, $f_2 = 0.210$)	0.813	0.806	0.469
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.28$)	0.137	0.247	0.837
0.05	0.1	Null	0.046 (0.37)	0.042 (0.37)	0.053 (0.21)
		DOM ($f_0 = 0.1$, $f_1 = f_2 = 0.336$)	0.804	0.780	0.336
		ADD ($f_0 = 0.1$, $f_1 = 0.332$, $f_2 = 0.564$)	0.815	0.810	0.672
		REC ($f_0 = f_1 = 0.1$, $f_2 = 0.58$)	0.148	0.230	0.824

[1]Adjusted; [2]unadjusted.

Control will be genotyping the several thousand participants in the National Health Interview Survey, in the future it should be possible to select null loci or find single-nucleotide polymorphisms (SNPs) whose frequencies are near that of the candidate allele. When this is feasible, $Z_0/\lambda_0^{1/2}$, $Z_{1/2}/\lambda_{1/2}^{1/2}$, or $Z_1/\lambda_1^{1/2}$ can be used when the underlying mode of inheritance is recessive, additive, or dominant, respectively. For a complex disease, however, the genetic model is usually unknown. Sometimes a previous segregation analysis would suggest the most plausible genetic model (Sham, 1998). This model might be tested by a procedure similar to that developed by Lee and Chang (2000) and Scherag et al. (2002) for the family-based trio studies. Then one can apply the GC test with the appropriate genetic model.

The GC test can be used to check an association found using a standard trend test. In this situation, the recessive (additive or dominant) GC test should be used when the trend test is optimal for the recessive (additive or dominant) model. For example, analyzing a case–control study, Tian et al. (2004) reported a significant association between Klippel–Trenaunay syndrome (KTS) and the angiogenic factor, VG5Q, when the mutation E133K of VG5Q is presented. The authors then applied the GC test with the additive scores to verify their finding. The association between KTS and

Table 4
Adjusted type I error of GC trend tests under different frequencies of null loci: $c = 41$ null loci and 100,000 simulations

	Gene frequency				
F	Candidate	Null loci	DOM	ADD	REC
		$a_1 = b_2 = 200$, $a_2 = b_1 = 0$			
0.005	0.1	(0.4, 0.6)	0.074	0.054	0.011
	0.1	(0.2, 0.4)	0.063	0.055	0.021
		$a_1 = a_2 = b_1 = b_2 = 100$			
	0.1	(0.4, 0.6)	0.057	0.059	0.054
	0.1	(0.2, 0.4)	0.057	0.058	0.053
		$a_1 = b_1 = 200$, $a_2 = b_2 = 0$			
	0.1	(0.4, 0.6)	0.057	0.059	0.052
	0.1	(0.2, 0.4)	0.057	0.058	0.051
		$a_1 = b_2 = 200$, $a_2 = b_1 = 0$			
0.05	0.1	(0.4, 0.6)	0.071	0.036	0.001
	0.1	(0.2, 0.4)	0.051	0.036	0.005
		$a_1 = a_2 = b_1 = b_2 = 100$			
	0.1	(0.4, 0.6)	0.058	0.068	0.048
	0.1	(0.2, 0.4)	0.059	0.067	0.046

VG5Q that they reported is based on a dominant genetic model. To confirm the result using the GC method, it would be appropriate for them to use the dominant GC test when null loci that whose frequencies close to that of E133K can be found.

It is not likely for a complex disease to strictly follow any of the three common genetic models. The fact that we have shown that the VIF depends on the allele frequency of the candidate gene for recessive and dominant models strongly suggests that it would be similarly dependent for other non-additive models. If all the genetic models are scientifically plausible or one cannot obtain null loci with frequencies near that of the candidate, then $Z_{1/2}/\lambda_{1/2}^{1/2}$, the additive GC test of Devlin and Roeder (1999), should be used as it is most robust across three genetic models and is independent of candidate gene frequency.

In this article, we considered the properties of GC tests for case–control studies under various genetic models. Other important methods and tests that account for the effect of population substructure, Pritchard and Donnelly (2001) and Satten, Flanders, and Yang (2001), have been developed for candidate gene association studies.

Table 5
Adjusted type I errors and power of GC trend tests under similar frequencies of null loci with two subpopulations ($a_1 = b_2 = 200$, $a_2 = b_1 = 0$): $c = 41$ null loci and 100,000 (type I error) and 10,000 (power) simulations

	Gene frequency					
F	Candidate	Null loci	Model	DOM	ADD	REC
0.005	0.1	(0.05, 0.15)	Null	0.055	0.056	0.037
			DOM	0.805	0.775	0.139
			ADD	0.794	0.786	0.324
			REC	0.176	0.311	0.788
	0.2	(0.1, 0.3)	Null	0.056	0.055	0.051
			DOM	0.781	0.706	0.203
			ADD	0.804	0.799	0.528
			REC	0.190	0.385	0.800
	0.5	(0.4, 0.6)	Null	0.055	0.056	0.056
			DOM	0.785	0.506	0.185
			ADD	0.781	0.788	0.680
			REC	0.228	0.555	0.766
0.05	0.1	(0.05, 0.15)	Null	0.049	0.047	0.061
	0.2	(0.1, 0.3)	Null	0.052	0.047	0.057

Acknowledgements

We thank the associate editor and two referees for their careful reviews, thoughtful comments, and helpful suggestions that greatly improved the manuscript. The research of Zhaohai Li was supported in part by grant EY14478 of The National Eye Institute and the work of Joseph Gastwirth was supported in part by grant SES-0317956 of The National Science Foundation.

Résumé

Les études cas-témoins sont fréquemment utilisées pour étudier l'association entre un allèle candidat et une maladie. Cependant, une association factice peut être observée du fait d'une stratification de population apparente ou cryptique qui augmente la variance du test de tendance. Devlin et Roeder ont dérivé ce facteur d'inflation de la variance (FIV) pour le test de tendance et ont proposé une nouvelle approche dite des 'contrôles génomiques' afin d'estimer le FIV et d'ajuster la statistique de test. Leurs résultats ont été dérivés sous l'hypothèse d'un modèle génétique additif de telle sorte que le FIV correspondant soit indépendant de la fréquence allélique de l'allèle candidat. Ici, nous déterminons le FIV appropriés pour des modèles récessif et dominant. Contrairement au cas additif, les FIV des tests optimaux pour ces deux modèles sont dépendants de la fréquence allélique de l'allèle candidat. Des études de simulation montrent que, lorsque les marqueurs nuls, i.e., les marqueurs utilisés pour estimer le FIV, ont des fréquences alléliques proches de celle du gène candidat, l'approche 'contrôles génomiques' reste optimale dans un contexte de modèle génétique dominant ou récessif. En revanche, lorsque le modèle génétique est inconnu ou lorsque les marqueurs nuls ont des fréquences alléliques différant de celles du gène candidat, l'approche 'contrôles génomiques' ne reste valable que pour un modèle génétique additif.

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11,** 375–386.

Bacanu, S.-A., Devlin, B., and Roeder, K. (2000). The power of genomic control. *American Journal of Human Genetics* **66,** 1933–1944.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* **10,** 417–451.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55,** 997–1004.

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology* **60,** 155–166.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sunderland, Massachusetts: Sinauer.

Freidlin, B., Zheng, G., Li, Z., and Gastwirth, J. L. (2002). Trend tests for case–control studies of genetic markers: Power, sample size and robustness. *Human Heredity* **53,** 146–152.

Lee, W. C. and Chang, C. H. (2000). Estimating genotype relative risks in case-parental control studies: An optimal weighting approach. *American Journal of Epidemiology* **152,** 487–492.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

Risch, N. and Teng, J. (1998). The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Research* **8,** 1273–1288.

Pritchard, J. K. and Donnelly, P. (2001). Case–control studies of association in structured or admixed populations. *Theoretical Population Biology* **60,** 227–237.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics* **53,** 1253–1261.

Satten, G. A., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case–control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics* **68,** 466–477.

Schaid, D. J. and Sommer, S. S. (1993). Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics* **53,** 1114–1126.

Scherag, A., Dempfle, A., Hinney, A., Hebebrand, J., and Schafer, H. (2002). Confidence intervals for genotype relative risks and allele frequencies from the case parent trio design for candidate-gene studies. *Human Heredity* **54,** 210–217.

Sham, P. (1998). *Statistics in Human Genetics*. London: Arnold.

Spielman, R. S., McGinnis, R., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52,** 506–516.

Tian, X.-L., Kadaba, R., You, S.-A., et al. (2004). Identification of an angiogenic factor that when mutated causes susceptibility to Klippel–Trenaunay syndrome. *Nature* **427,** 640–645.

Zheng, G., Freidlin, B., Li, Z., and Gastwirth, J. L. (2003). Choice of scores in trend tests for case–control studies of candidate-gene associations. *Biometrical Journal* **45,** 335–348.

Received December 2003. *Revised April* 2004.
Accepted June 2004.

Appendix A

There are 16 possible pairs of genotypes for two alleles, e.g., $AA \times AA$, $Aa \times AA$, etc. Twelve of them are given in Table 2. We derive the joint distribution of the pair $(AA,\ AA)$. Let $A = A$, $B = A$, $C = A$, and $D = A$. In the following $A \equiv B$ is used to denote that A and B are IBD, etc. The following events are mutually exclusive: all four alleles are IBD ($A \equiv B \equiv C \equiv D$), three alleles are IBD ($A \equiv B \equiv C$, $A \equiv B \equiv D$, $A \equiv C \equiv D$, or $B \equiv C \equiv D$), two pairs of alleles are IBD ($A \equiv B$ and $C \equiv D$, $A \equiv C$ and $B \equiv D$, or $A \equiv D$ and $B \equiv C$), two alleles are IBD ($A \equiv B$, $A \equiv C$, $A \equiv D$, $B \equiv C$, $B \equiv D$, or $C \equiv D$), and no alleles are IBD. By Law of Total Probability, $\Pr(AA,\ AA) = \delta_4 p + (4\delta_3 + 3\delta_{22})p^2 + 6\delta_2 p^3 + \delta_0 p^4$. Substituting the five probabilities δ_0, δ_2, δ_3, δ_{22}, δ_4 as functions of F yields the first entry in Table 2.

Appendix B

For any R and $S, T^{\text{rec}} = S\sum_i G_i^{\text{rec}} - R\sum_j H_j^{\text{rec}}$. "rec" is not shown below.

$$\begin{aligned}
\text{var}(T) &= S^2 \sum_{i=1}^{R} \text{var}(G_i) + R^2 \sum_{j=1}^{S} \text{var}(H_j) + 2S^2 \sum_{i<l} \text{cov}(G_i, G_l) \\
&\quad + 2R^2 \sum_{j<l} \text{cov}(H_j, H_l) - 2RS \sum_i \sum_j \text{cov}(G_i, H_j) \\
&= RS(R+S)\text{var}(G_1) \\
&\quad + \sum_{k=1}^{m} \left\{ a_k(a_k-1)S^2 + b_k(b_k-1)R^2 - 2a_k b_k RS \right\} \\
&\quad \times \text{cov}(G_1, G_2),
\end{aligned}$$

where $\text{var}(G_1) = pq(p+Fq)\{1+(1-F)p\}$, $\sum_k \{a_k(a_k-1)S^2 + b_k(b_k-1)R^2 - 2a_k b_k RS\} = \sum_k (a_k S - b_k R)^2 - RS(R+S)$, and $\text{cov}(G_i, G_l)$ is given by (3). Thus, we obtain

$$\begin{aligned}
\text{var}(T) &= RS(R+S)pq(p+Fq)\{1+(1-F)p\} \\
&\quad - RS(R+S)\frac{2pqF(2p^2+p(5-3p)F+(3-4p)F^2-pqF^3)}{(1+F)(1+2F)} \\
&\quad + \frac{2pqF(2p^2+p(5-3p)F+(3-4p)F^2-pqF^3)}{(1+F)(1+2F)} \\
&\quad \times \sum_k (a_k S - b_k R)^2 \\
&= \frac{RS(R+S)p^2q(1+p)}{(1+F)(1+2F)} + \frac{RS(R+S)pqF}{(1+F)(1+2F)} h_1(F,p) \\
&\quad + \frac{2pqF}{(1+F)(1+2F)} h_2(F,p) \sum_k (a_k S - b_k R)^2.
\end{aligned}$$